Modelo scoring para recuperar cartera de microcredito

Presentado por
Angelica Lucia Rodriguez Avellaneda

Fundación Universitaria Los Libertadores
Facultad de Ingeniería y Ciencias Básicas
Especialización en Estadística Aplicada
Bogotá D.C, Colombia
2018

Modelo scoring para recuperar cartera de microcredito

Presentado por

Angelica Lucia Rodriguez Avellaneda

en cumplimiento parcial de los requerimientos para optar al título de

Especialista en Estadística Aplicada

Dirigida por

Juan Camilo Santana

Profesor

Fundación Universitaria Los Libertadores

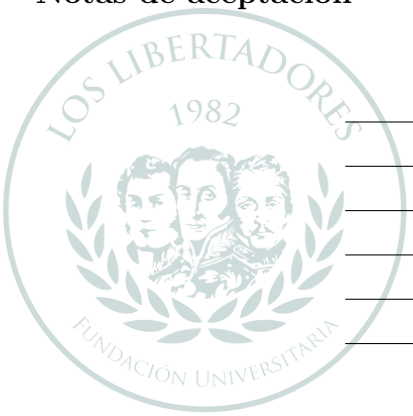
Facultad de Ingeniería y Ciencias Básicas

Especialización en Estadística Aplicada

Bogotá D.C, Colombia

2018

Notas de aceptación



Firma del presidente del jurado

Firma del jurado

Firma del jurado

Bogotá DC, Diciembre de 2018.

LOS LIBERTADORES

FUNDACIÓN UNIVERSITARIA

Las directivas de la Fundación Universitaria Los Libertadores, los jurados calificadores y el cuerpo docente no son responsables por los criterios e ideas expuestas en el presente documento. Estos corresponden únicamente a los autores y a los resultados de su trabajo.

Dedicatoria

Este documento se lo dedico a todas las personas que han contribuido en mi desarrollo profesional, especialmente a mi familia, los cuales me han apoyado incondicionalmente.

Agradecimientos

Gracias a la universidad y a la empresa en donde trabajo por darme la motivación de quererme exigirme más, y brindarme las herramientas necesarias para hacer de este trabajo una realidad.

Índice general

1 Introducción **3**

2 Planteamiento del Problema **5**
2.1 Objetivos 6
2.1.1 Objetivo General 6
2.1.2 Objetivos Específicos 6
2.2 Justificación 7

3 Marco Teórico / conceptual **9**
3.1 Microcredito 9
3.1.1 Caracteristicas 9
3.1.2 Cliente Micro-crediticio 9
3.2 Recuperación de los clientes 10
3.3 Variables 10
3.3.1 Perfil del cliente: 10
3.3.2 Comportamiento histórico: 11
3.4 Modelos Scoring 11
3.4.1 Modelo regresión logística 11
3.4.2 Modelo regresión probit 12
3.4.3 Redes Neuronales 12
3.4.4 Arboles de decisión 13

4 Marco Metodológico **15**
4.1 Segmentación Base 16
4.2 Modelos 16

5 Análisis y Resultados **17**
5.1 Regresión Logística 17
5.1.1 Modelo Inicial 17
5.1.2 Método akaike 17
5.1.3 Modelo Final 19

5.1.4 Matriz de confusión 19
5.2 Regresión Probit 20
5.2.1 Modelo Inicial 20
5.2.2 Método akaike 21
5.2.3 Modelo Final 22
5.2.4 Matriz de confusión 22
5.3 Redes Neuronales 22
5.3.1 Matriz de Confusión 23
5.4 Arboles de decisión 24
5.4.1 Matriz de Confusión 24
5.5 Relación de las variables con la mora del cliente 24

6 Conclusiones y Recomendaciones **29**

Índice de figuras

2.1 Comparativo indicador calidad de cartera. 5

3.1 Calificación de los clientes 10
3.2 Redes Neuronales 13
3.3 Arboles de decisión 14

Índice de cuadros

5.1 Modelo Inicial Regresión Logística 17

5.2 Método Akaike Regresión Logística 19

5.3 Modelo Final Regresión Logística 19

5.4 Matriz de confusión regresión logística base Entrenamiento. 20

5.5 Matriz de confusión regresión logística base Testeo. 20

5.6 Modelo Inicial Regresión Probit 21

5.7 Método Akaike Regresión Probit 21

5.8 Modelo Final Regresión Probit 22

5.9 Matriz de confusión regresión probit base Entrenamiento. 22

5.10 Matriz de confusión regresión probit base Testeo. 23

5.11 Matriz de confusión Redes Neuronales base Entrenamiento. 23

5.12 Matriz de confusión Redes Neuronales base Testeo. 23

5.13 Matriz de confusión Arboles de decisión base Entrenamiento. 24

5.14 Matriz de confusión Arboles de decisión base Testeo. 24

5.15 Relación variables con la mora 25

Modelo scoring para recuperar cartera de microcredito

Resumen

En el presente trabajo se expone la construcción de un modelo Scoring para determinar la probabilidad de pago del cliente cuando se encuentra en mora, enfocado al sector micro crediticio, buscando una herramienta con la cual se pueda gestionar la cartera de una manera eficiente, reduciendo los costos e incrementando el nivel de recuperación. A fin de cumplir este objetivo se cuenta con una base de 6 meses (abril- septiembre del 2018), con la cual se generan 4 modelos diferentes, los cuales son: regresión logística, regresión probit, arboles de decisión y redes neuronales. Se compara el porcentaje de error en cada uno de ellos, dando como resultado que el porcentaje menor el generado por medio de árboles de decisión con un 11.3 %, dicho modelo refleja un ajuste adecuado, mostrando un nivel de predictibilidad acertado. En cuanto a los modelos restantes ninguno cuenta con un porcentaje mayor al 12.5 %. Por lo cual por medio de la regresión logística se determina que la garantía, el género y la mora al momento de evaluación son las variables con mayor incidencia en el no pago de los clientes.

Palabras claves: Probabilidad de pago, scoring, microcrédito, arboles de decisión.

Abstract

In the present work the construction of a Scoring model is exposed to determine the probability of payment of the client when it is in default, focused on the micro credit sector, looking for a tool with which the portfolio can be managed in an efficient way, reducing the Costs and increasing the level of recovery. In order to meet this objective, we have a base of 6 months (April-September 2018), which presents 4 different models, which are: logistic regression, probit regression, decision trees and neural networks. The percentage of error in each of them is compared, the result is an appropriate percentage of error. As for the remaining models, none has a percentage greater than 12.5 %. Therefore, through logistic regression it is determined that the guarantee, the gender and the default at the time of evaluation are the variables with the highest incidence in the non-payment of the clients.

Keywords: Probability of payment, punctuation, microcredit, decision trees.

Capítulo 1

Introducción

El presenta trabajo se encuentra estructurado en 5 capítulos adicionales a este. El siguiente indica el planteamiento del problema a tratar el cual describe la importancia de realizar un modelo para determinar la probabilidad de pago de los clientes en el sector micro crediticio dado que dicho sector al contar con un perfil de clientes con un nivel de riesgo más alto, el indicador de mora incrementa significativamente respecto a los demás sectores.

En el tercero se realiza una revisión de conceptos identificando las características del sector microcreditico, y adicionalmente se realiza una revisión bibliográfica con el fin de establecer las variables que harán parte del modelo, las cuales se dividen en dos grupos perfil del cliente por ejemplo la edad, el género y el nivel de ingresos y su comportamiento histórico como mora máxima en los últimos 6 meses.

El cuarto capítulo expone la metodología utilizada, en la cual se explica que se realizaron 4 modelos (regresión logística, regresión probit, arboles de decisión y redes neuronales) por medio del aplicativo R, adicionalmente se menciona que con el fin de identificar la efectividad del modelo la base fue divida en dos partes homogéneas, con una se genera los modelos y con la otra se valida el porcentaje e error, para lo cual se cuenta con una base de 6 meses de una entidad micro crediticia.

El quinto capítulo corresponde al análisis y resultados en donde se expone la matriz de confusión de cada modelo, en donde arboles de decisiones cuenta con el menor porcentaje de error (11.3 %), asimismo se identifican las variables que afectan el incremento de la mora.

Por último, las conclusiones, en donde se menciona que el modelo presenta una discriminación adecuada de la probabilidad de pago en los clientes, por lo cual puede ser de gran utilidad en el sector micro crediticio para la aplicación de estrategias diferenciales.

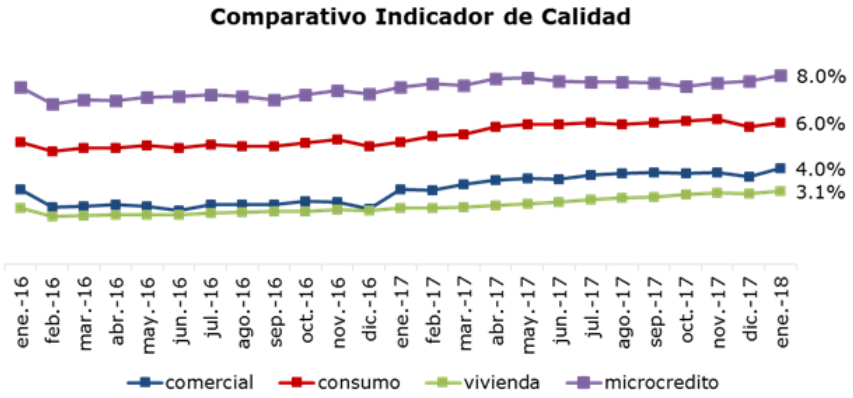
Capítulo 2

Planteamiento del Problema

¿Cómo diferenciar los clientes que se encuentran en mora por medio de su probabilidad de pago en el sector micro crediticio?

En Colombia existen bancos dedicados al sector micro crediticio, generando alternativas diferenciales a personas que cuentan con algún tipo de negocio y en su mayoría, son de escasos recursos, por lo cual no cuentan con una vida crediticia significativa, dificultando el ingreso a la banca tradicional. Al tener este perfil de clientes el nivel de riesgo es más alto, comparado con otros tipos de cartera, ocasionando que el índice de calidad de cartera sea mayor.

Figura 2.1: Comparativo indicador calidad de cartera.



Fuente:[1].

Como se puede observar en la gráfica, de acuerdo a la superfinanciera a corte enero el indicador de calidad (saldo capital en mora mayor a 30 días / saldo capital total) se encuentra en el 8 % mientras el sector de vivienda, consumo, y comercio se encuentra en el 3.1 %, 6 % y 4 % respectivamente. Lo cual ocasiona que el esfuerzo en términos de recuperación sea significativo. Sin embargo, en su mayoría, comparado con los bancos tradicionales este tipo de entidades no cuentan con la tecnología ni modelos estadísticos

estructurados que ayuden a identificar el perfil de cada cliente, es medir, modelos en los cuales se determine la probabilidad de que el cliente realice el pago de su obligación.

2.1 Objetivos

2.1.1 Objetivo General

Construir un modelo estadístico que permita cuantificar la probabilidad de pago de clientes en mora sobre una cartera de microcredito

2.1.2 Objetivos Específicos

- Comparar diferentes metodologias, que permitan evaluar las respuestas de un cliente que esta en mora.
- Identificar las variables que afectan el incremento de la mora de los clientes.

2.2 Justificación

De acuerdo a la revisión bibliográfica realizada, se identifica que los estudios se encuentran ligados principalmente al otorgamiento, es decir si es o no recomendable concederle el crédito a la persona, a continuación algunos ejemplos:

"El modelo logístico: Una herramienta estadística para evaluar el riesgo de crédito", elaborado en la universidad de Medellín en donde el objetivo era determinar la viabilidad de otorgar un crédito, por medio de la regresión logística, el modelo presenta un buen ajuste con un margen de error inferior al 12 %. *Fuente:[2].*

"La ejecución de un modelo de implementación de un Scoring estadístico para las micro finanzas" en la Universidad Nacional de Trujillo, donde el objetivo del paper era determinar la viabilidad de otorgar un crédito para la cartera micro crediticia, por medio del modelo realizado se incrementa el 19 % la predictibilidad respecto al utilizado anteriormente.Fuente:[3]

En cuanto a modelos que permitan determinar la probabilidad de pago del cliente, principalmente se observan en el sector consumo y comercial, donde por ejemplo en Ecuador, en el estudio denominado "Desarrollo de un modelo de scoring de segmentación de cobranzas para tarjeta de crédito de la banca de personas de Produbanco", se indicaba que el modelo con un porcentaje de error inferior al 10 %, cumple con el objetivo de obtener la probabilidad de mora, con la cual se realiza la segmentación de gestión de cobro, identificando las variables que son influyentes al momento de determinar el riesgo de no pago. Fuente:[4]

Sin embargo para microcrédito no se observan este tipo de modelos, por lo cual, dado el índice de mora explicado previamente y la composición de esta cartera, surge la necesidad de realizar un modelo que presente una capacidad de discriminación entre clientes buenos y malos, identificando si las bondades descritas anteriormente se pueden observar en dicho sector, con el propósito de establecer estrategias focalizadas evitando utilizar los recursos del banco de forma incorrecta, e incrementando los niveles de recuperación.

Capítulo 3

Marco Teórico / conceptual

3.1 Microcredito

Son préstamos cuyo monto es pequeño (promedio de $4.6 Millones) y cuenta con condiciones especiales de tasa de interés y plazos de amortización, su objetivo principal es financiar proyectos que ya están en marcha o para el mejoramiento de producción de microempresas (inversiones en activos fijos o capital de trabajo, entre otros).

3.1.1 Caracteristicas

- Mecanismo de financiación para empresas formales e informales.
- La amortización o pago del capital depende de cada proyecto y puede ser de corto o mediano plazo.
- Montos de financiación que se ajustan a las necesidades de la empresa.
- Dependiendo de las políticas de la entidad se pueden asignar a un si el solicitante no ha tenido experiencia previa en el sector financiero. Fuente:[5]

3.1.2 Cliente Micro-crediticio

- Personas que cuentan con una unidad de negocio, el cual no necesariamente debe ser una empresa.
- Personas sin historial crediticio, los cuales difícilmente tienen acceso a la banca tradicional.
- Personas de escasos recursos, tanto en el sector urbano como en el sector rural.

3.2 Recuperación de los clientes

De acuerdo a la Superintendencia financiera cada cliente es calificado de acuerdo a su mora, y así mismo se debe provisionar un valor. Existen 5 calificaciones, A, B, C, D y E para el sector micro crediticio funciona de la siguiente manera:

Figura 3.1: Calificación de los clientes

Categoría	Descripción	N° Meses en Mora
A: "Riesgo Normal"	El cliente cuenta con una capacidad de pago adecuada, en términos del monto y origen de los ingresos.	Hasta 1 mes
B: "Riesgo Aceptable"	Los créditos están aceptablemente atendidos y protegidos, pero existen debilidades que potencialmente pueden afectar, la capacidad de pago del deudor.	Entre 1 y 2 meses
C: "Crédito deficiente con riesgo apreciable"	Los créditos presentan insuficiencias en la capacidad de pago del deudor y comprometen el normal recaudo de la obligación.	Entre 2 y 3 meses
D: "Crédito de difícil cobro con riesgo significativo"	Su probabilidad de recaudo es altamente sudosa	Entre 3 y 4 meses
E: "Riesgo de incobrabilidad"	Se estima incobrable	Mayor a 4 meses

Fuente:[5]

Por lo cual, se define:

Clientes Recuperados: Créditos que finalizan el mes, con mora menor a los 30 días, dado que se encuentran catalogados en la categoría A.

Clientes no Recuperados: Créditos que finalizan el mes, con mora mayor a los 30 días, correspondiente a las categorías restantes.

3.3 Variables

De acuerdo a los estudios realizados las variables se dividen en dos grandes grupos:

3.3.1 Perfil del cliente:

Hace referencia a la información socio demográfico del cliente e información ingresada al momento de aprobación del crédito

- Edad
- Monto desembolsado
- Antigüedad del crédito en meses

- Nivel ingresos
- Total activos
- Garantía: servicios financiero que respaldan las obligaciones contraídas con la entidad.
- Sector: si es Urbano o Rural
- Sector agropecuario: Actualmente se está fortaleciendo el sector incrementando su financiamiento, por ejemplo FINAGRO creo líneas de microcrédito en zonas rurales para población de bajos recursos, incluso en pobreza extrema a clientes catalogados como productores

3.3.2 Comportamiento histórico:

- Porcentaje cuotas pagadas
- Mora fecha evaluación
- Maxima mora en los últimos 6 meses
- Veces mora (No veces que el cliente se encontró en mora mayor a 30 días durante los últimos 6 meses)
- Pagos (No de pagos realizados en los últimos 6 meses)

3.4 Modelos Scoring

Permiten a las organizaciones transformar los datos crediticios para predecir el comportamiento futuro de sus clientes, lo cual conlleva a tomar decisiones con mayor rapidez y limitar las pérdidas mediante una gestión integral, dado que es posible identificar el perfil de los clientes, es decir si son "buenos" o "malos" y así establecer los próximos pasos, tanto en términos de colocación como en términos de recuperación.

De acuerdo a la revisión bibliográfica realizada unos de los métodos estadísticos más utilizados para el diseño de modelos de Scoring son: Modelo regresión logística, Arboles de decisión y redes Neuronales. Por lo cual a continuación se muestra información sobre estos modelos Fuente:[6]

3.4.1 Modelo regresión logística

Su principal objetivo es estudiar el efecto de variables explicativas sobre una variable dicotómica, es decir que toma únicamente dos valores (0 o 1), estableciendo la probabilidad de ocurrencia de un evento teniendo en cuenta diferentes factores. Las variables explicativas pueden ser cuantitativas y cualitativas tomando una función exponencial. Adicionalmente, de acuerdo al modelo generado es posible tener que utilizar las variables Dummies. Fuente:[7]

Este modelo tiene forma:

$$\mathrm{y}_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

para i=1,2,...,n

De esto se traduce que:

$$\text{y=1, } \varepsilon_i = 1 - \beta_0 - \beta_1 x_i$$

$$\text{y=0, } \varepsilon_i = -\beta_0 - \beta_1 x_i$$

- Donde Yi: Representa la variable dependiente
- Bj: Representa los parámetros.
- Xi: Variable independiente

La probabilidad que ocurra un suceso dado que el individuo presenta los valores X1, X2, Xp se muestra de la siguiente manera

$$\text{Pr(Y=1—x}_1, x_2, ..., x_p) = \frac{1}{1+exp(-\alpha-\beta_1 x_1-\beta_2 x_2-...-\beta_p x_p)}$$

Es por tanto, una técnica multivariante de dependencia ya que trata de estimar la probabilidad de que ocurra un suceso en función de la dependencia de otras variables. *Fuente:[8]*

3.4.2 Modelo regresión probit

De la misma manera que la regresión logística, la variable dependiente es dicotómica es decir toma valores de 0 y 1, la diferencia con la regresión logística es que se utiliza la función de distribución normal estándar evaluada de la siguiente manera:

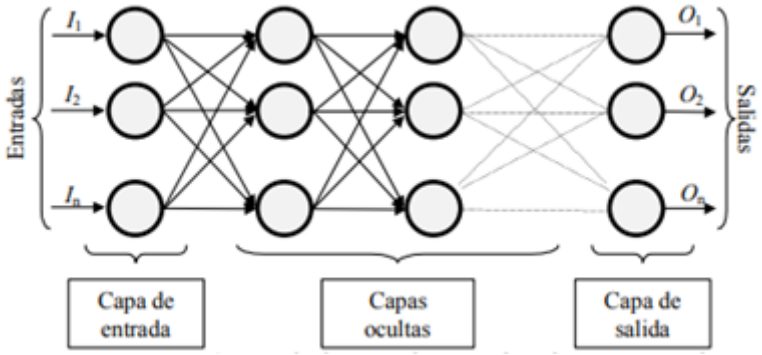
$$\text{Z=}\beta_0 + \beta_1 X : Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1 X)$$

Donde: B es el "valor z"ó "índice z"del modelo probit *Fuente:[9]*

3.4.3 Redes Neuronales

Es un modelo matemático inspirado en el comportamiento biológico de las neuronas y en cómo se organizan formando la estructura del cerebro, dado que las redes se encuentran constituidas por neuronas interconectadas y arregladas en diferentes capas de la siguiente manera:

Figura 3.2: Redes Neuronales



Fuente:[10]

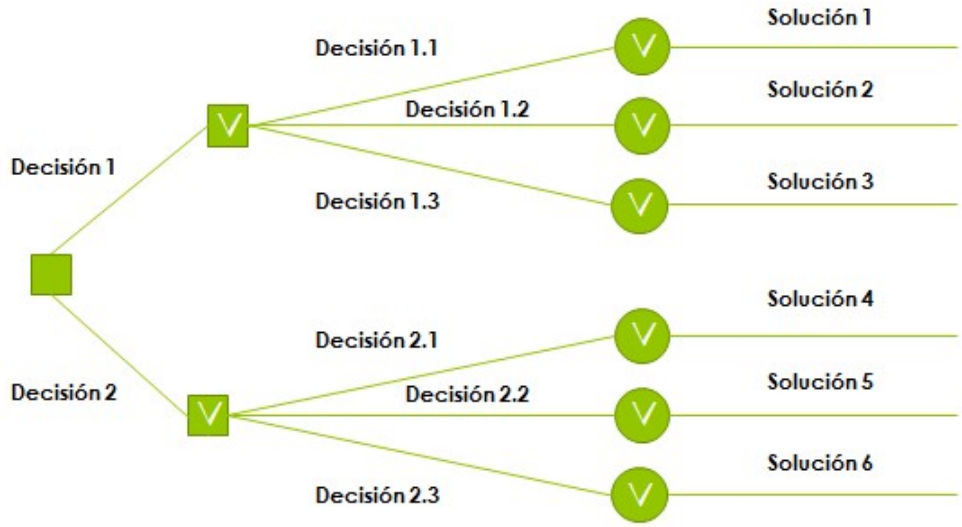
Los datos ingresan por medio de la capa de entrada, pasan a través de la capa oculta los valores se propagan desde cada neurona hasta cada neurona de la capa siguiente, en esta capa se estructura el modelo por lo cual pueden ser diferentes capas, y se denomina de esta manera porque la red se comporta como una "caja negra" es decir no es posible observar el procedimiento, únicamente los resultados, los cuales se muestran en la capa de salida *Fuente:[11]*

La red aprende a través del entrenamiento examinando los registros individuales, generando una predicción para cada registro y realizando ajustes a las ponderaciones cuando realiza una predicción incorrecta. Este proceso se repite muchas veces y la red sigue mejorando sus predicciones hasta haber ajustado el mejor modelo. *Fuente:[12]*

3.4.4 Arboles de decisión

Un árbol de decisión es una forma gráfica facilita el análisis y la toma de decisiones dado que permite visualizar las interacciones entre las rutas y las probabilidades de ocurrencia de los eventos, permitiendo identificar el mejor modelo probabilístico, como se puede observr en la siguiente grafica:. *Fuente:[13]*

Figura 3.3: Arboles de decisión



Fuente:[14]

El algoritmo utilizado en este trabajo es el C5.0 el cual fue desarrollado por Ross Quinlan y examina todos los campos del conjunto de datos para detectar el que proporciona la mejor clasificación, divide los datos en subgrupos y asi mismo los subgrupos se dividen en unidades cada vez más pequeñas hasta completar el árbol. *Fuente:[15]*

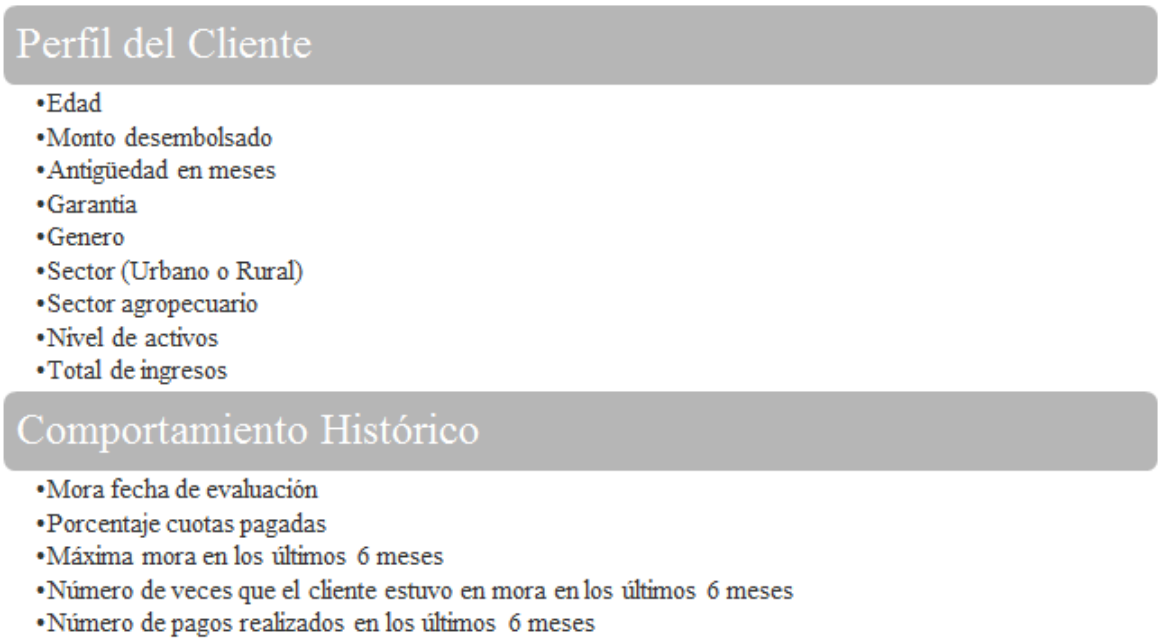
Capítulo 4

Marco Metodológico

La entidad financiera permite el acceso a la información requerida para el desarrollo del modelo, con la condición de no entregar la base de datos a ningún tercero, y no mencionar el nombre de la empresa. Para extraer la información se utiliza el programa SQL el cual permite el acceso y manipulación de información en una base de datos donde se encuentra la información demográfica y de su comportamiento histórico, permitiendo seleccionar las variables para este modelo.

La base de datos utilizada está constituida por 109.000 registros de clientes con mora entre 30 y 90 días en los meses de abril a septiembre.

La selección de las variables se realiza de acuerdo a los estudios realizados previamente:



Elaboración propia

La variable dependiente es una variable dummy en donde se indica si el cliente al finalizar el mes se recuperó, es decir si su mora finalizo en menos de 30 días.

Debido a que las variables garantía, género, sector y sector agropecuario son variables categóricas, para los modelos toman valores de 0 y 1 de la siguiente manera:

Variables	0	1
Garantía	No tiene garantía	Tiene garantía
Género	Femenino	Masculino
Sector	Rural	Urbano
Sector agropecuario	No agro	Agropecuario

4.1 Segmentación Base

Con el fin de identificar la efectividad del modelo la base fue divida en dos partes homogéneas (50 % cada una), de la siguiente manera:

• Base entrenamiento: A partir de esta base se genera cada uno de los modelos

• Base para realizar testeo: Es utilizada para probar la efectividad de los modelos generados, determinando el porcentaje de error.

4.2 Modelos

Por medio del aplicativo R se generan 4 modelos diferentes, determinando cual tiene un porcentaje de error menor, es decir una mejor efectividad, dichos modelos fueron:

• Arboles de decisión: Se realizan por medio de la función C5.0 la cual ajusta el modelo basado en el algoritmo desarrollado por Ross Quinlan

• Redes Neuronales: Por medio de la función nnet la cual ajusta la red con una capa oculta

• Regresión logística: Por medio de la función glm, adicional se utiliza el método de Akaike, determinando cuales son las variables significativas en el modelo.

• Regresión probit: Por medio de la función gml indicando que el modelo a generar pertenece a la familia binomial probit, adicional de la misma manera que en la regresión logística, se utiliza el método de Akaike, determinando cuales son las variables significativas en el modelo

Capítulo 5

Análisis y Resultados

5.1 Regresión Logística

5.1.1 Modelo Inicial

Como modelo inicial se toman todas las variables descritas anteriormente, sin embargo se observa que no todas son significativas, teniendo en cuenta un nivel de significancia del 5 %:

Cuadro 5.1: Modelo Inicial Regresión Logística

Variables	Pr ($\|z\|$)	
intercepto	<2e-16	***
Mora inicial	<2e-16	***
antigüedad	<2e-16	***
Monto	0,0201	*
Garantia	<2e-16	***
Genero	0,0003	***
Sector	0,441	
Nivel ingresos	0,877	
Total Activos	0,469	
Edad	0,018	*
sector Agropecuario	0,023	*
porcentaje pago : N pagos	8,9e-7	***
max mora : N veces en mora	<2e-16	***

5.1.2 Método akaike

Dado lo anterior se genera el método akaike, el cual es una medida de bondad de ajuste en un modelo estadístico indicando que el modelo más adecuado con un AIC de 34.339

es:

glm(formula= Recuperado mora_inicial + antigüedad + monto + garantía + genero + edad + agropecuario + P_pago : N_pagos + max_mora : veces_mora, family = binomial)

Cuadro 5.2: Método Akaike Regresión Logística

Variable	Coeficiente
Intercepto	2,844
Mora_inicial	-0,620
Antigüedad	0,618
Monto	-0,007
Garantia	-0,409
Genero	-0,103
Edad	0,002
agropecuario	0,075
P_pago : N_pagos	-0,065
max_mora : veces_mora	0,002

5.1.3 Modelo Final

Se genera nuevamente el modelo con las variables arrojadas en el método akaike, identificando que todas las variables son significativas, eliminando el sector, nivel de ingresos, y los activos del cliente

Cuadro 5.3: Modelo Final Regresión Logística

Variables	Pr ($\|z\|$)	
intercepto	<2e-16	***
Mora inicial	<2e-16	***
antigüedad	<2e-16	***
Garantia	<2e-16	***
Genero	0,0003	***
Monto	0,022	*
Edad	0,018	*
sector Agropecuario	0,023	*
porcentaje pago : N pagos	7,94e-7	***
max mora : N veces en mora	<2e-16	***

5.1.4 Matriz de confusión

Por medio del modelo anteriormente descrito se determina el porcentaje de error del modelo, es decir los clientes que califica de manera incorrecta.

Cuadro 5.4: Matriz de confusión regresión logística base Entrenamiento.

	0	1	Total
0	18.126	4.638	22.764
1	2.190	29.754	31.944
Total	20.316	34.392	54.708
% Error		6.828	12,5 %

Como se puede observar en el cuadro anterior el porcentaje de error es del 12.5 % con 6.828 registros erroneos de 54.708

Cuadro 5.5: Matriz de confusión regresión logística base Testeo.

	0	1	Total
0	19.445	3.330	22.775
1	3.186	28.667	31.853
Total	22.631	31.997	54.628
% Error		6.516	11,9 %

Para la base que no se tuvo en cuenta en la construcción del modelo el porcentaje de error disminuye al 11,9 %, con 6.516 registros erroneos de 54.628, de los cuales 3.330 fueron calificados como recuperados y 3.186 como no recuperados cuando la situación es contraria. Dado que el porcentaje se encuentra alrededor del 12 % se resalta que el modelo cuenta con un buen ajuste y una predictibilidad acertada.

5.2 Regresión Probit

5.2.1 Modelo Inicial

Como modelo inicial se toman todas las variables descritas anteriormente, sin embargo como sucedió para la regresión logistica se observa que no todas son significativas, teniendo en cuenta un nivel de significancia del 5 %:

Cuadro 5.6: Modelo Inicial Regresión Probit

Variables	Pr ($\|z\|$)	
intercepto	<2e-16	***
Mora inicial	<2e-16	***
antigüedad	<2e-16	***
Monto	0,059	.
Garantia	<2e-16	***
Genero	0,0002	***
Sector	0,501	
Nivel ingresos	0,876	
Total Activos	0,487	
Edad	0,010	*
sector Agropecuario	0,028	*
porcentaje pago : N pagos	1,88e-6	***
max mora : N veces en mora	<2e-16	***

5.2.2 Método akaike

Por lo cual se genera el método akaike, de la misma manera que en el modelo anterior, con un AIC de 34.563 se obtiene el siguiente modelo:

glm(formula= Recuperado mora_inicial + antigüedad + monto + garantía + genero + edad + agropecuario + P_pago : N_pagos + max_mora : veces_mora, family = binomial(link="probit")

Cuadro 5.7: Método Akaike Regresión Probit

Variable	Coeficiente
Intercepto	1,653
Mora_inicial	-0,034
Antigüedad	0,316
Monto	-0,003
Garantia	-0,232
Genero	-0,056
Edad	0,0014
agropecuario	0,040
P_pago : N_pagos	-0,033
max_mora : veces_mora	0,001

5.2.3 Modelo Final

Se genera nuevamente el modelo en donde se encuentra que todas las variables son significativas, eliminando las mismas variables que en regresión logística

Cuadro 5.8: Modelo Final Regresión Probit

Variables	Pr ($\|z\|$)	
intercepto	<2e-16	***
Mora inicial	<2e-16	***
antigüedad	<2e-16	***
Garantia	<2e-16	***
Genero	0,0002	***
Monto	0,044	*
Edad	0,011	*
sector Agropecuario	0,022	*
porcentaje pago : N pagos	1,98e-6	***
max mora : N veces en mora	<2e-16	***

5.2.4 Matriz de confusión

Por medio del modelo anteriormente descrito se determina el porcentaje de error del modelo, es decir los clientes que califica de manera incorrecta.

Cuadro 5.9: Matriz de confusión regresión probit base Entrenamiento.

	0	1	Total
0	17.812	4.952	22.764
1	2.002	29.942	31.944
Total	19.814	34.894	54.708
% Error		6.954	12,7 %

Como se puede observar en el cuadro anterior el porcentaje de error es del 12.7 % con 6.954 registros erroneos de 54.708, es decir 23Pbs por encima de la regresión logística.

Para la base de testeo el porcentaje de error disminuye al 12,3 %, con 6.708 registros erroneos de 54.628, este porcentaje es 43Pbs superior al generado en el modelo de regresión logistica.

5.3 Redes Neuronales

Para este modelo únicamente se puede observar el resultado final, es decir el que cuenta con el mejor ajuste, dado que de acuerdo a lo mencionado anteriormente el modelo se

Cuadro 5.10: Matriz de confusión regresión probit base Testeo.

	0	1	Total
0	20.066	2.709	22.775
1	3.999	27.858	31.853
Total	24.065	30.567	54.628
% Error		6.708	12,3 %

estructura en la capa oculta la cual se comporta como una çaja negra".

De la misma manera que en los modelos descritos anteriormente, se realiza la matriz de confusión para las dos bases.

5.3.1 Matriz de Confusión

Cuadro 5.11: Matriz de confusión Redes Neuronales base Entrenamiento.

	0	1	Total
0	19.190	3.574	22.764
1	2.899	29.045	31.944
Total	22.089	32.619	54.708
% Error		6.473	11,8 %

Se observa en el cuadro anterior que el porcentaje de error es del 11.8 % con 6.473 registros erroneos de 54.708, 65Pbs por debajo comparado con el modelo de regresión logistica, por lo cual hasta el momento este modelo presenta el porcentaje de efectividad más alto.

Cuadro 5.12: Matriz de confusión Redes Neuronales base Testeo.

	0	1	Total
0	19.248	3.527	22.775
1	2.905	28.948	31.853
Total	22.153	32.475	54.628
% Error		6.432	118 %

Para la base de testeo porcentaje de error disminuye ligeramente pasando del 11,9 % al 11,8 %, con 6.432 registros erroneos de 54.628, en conclusión aunque en la base de entrenamiento se presentaba una diferencia significativa con los dos modelos anteriores, en la base de testeo esta diferencia disminuye a unicamente 15Pbs.

5.4 Arboles de decisión

Por medio del algoritmo C5.0 el aplicativo R determina el mejor modelo. Continuando con la metodología anteriormente descrita, se realiza la matriz de confusión para las dos bases.

5.4.1 Matriz de Confusión

Cuadro 5.13: Matriz de confusión Arboles de decisión base Entrenamiento.

	0	1	Total
0	19.719	2.906	22.625
1	3.045	28.138	31.183
Total	22.764	31.044	53.808
% Error		5.951	11,1 %

Se observa en el cuadro anterior que el porcentaje de error es del 11,1 % con 5.951 registros erroneos de 53.808, 77Pbs por debajo comparado con el modelo de redes neuronales, por lo cual de los 4 modelos realizados, este es el que presenta el porcentaje de efectividad más alto.

Cuadro 5.14: Matriz de confusión Arboles de decisión base Testeo.

	0	1	Total
0	19.685	3.073	22.758
1	3.090	28.780	31.870
Total	22.775	31.853	54.628
% Error		6.163	11,3 %

Para la base de testeo porcentaje de error incrementa ligeramente pasando del 11.1 % al 11,3 %, con 6.163 registros erroneos de 54.628, por lo tanto de acuerdo a lo observado en la base de entrenamiento el modelo realizado por medio de los arboles de decisión refleja el mejor ajuste y por ende la acertividad más alta.

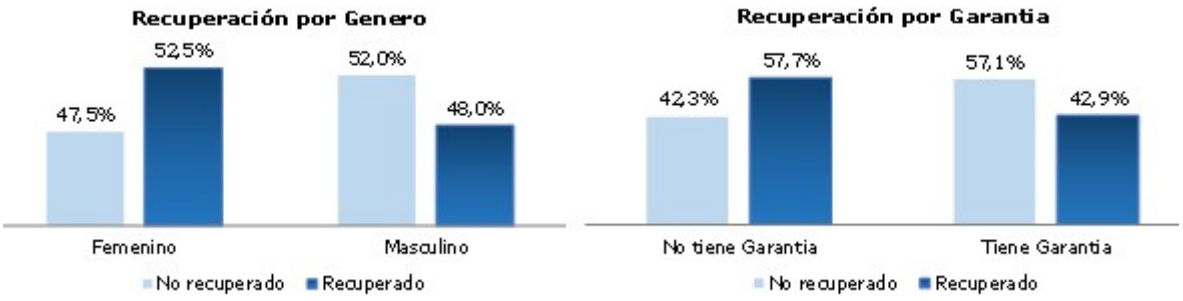
5.5 Relación de las variables con la mora del cliente

Este indicador muestra si la variable afecta la mora del cliente, entre más bajo su incidencia es mayor, por esta razón la garantía, el género y la mora al momento de generar el modelo son las variables que más afectan este comportamiento, dicha relación se determina por medio de la regresión logística.

Cuadro 5.15: Relación variables con la mora

Variable	Coeficiente
Intercepto	17,199
Mora_inicial	0,939
Antigüedad	1,856
Monto	0,992
Garantia	0,663
Genero	0,902
Edad	1,002
agropecuario	1,078
P_pago : N_pagos	0,936
max_mora : veces_mora	1,002

Dado lo anterior se analizan las variables, con el fin de establecer el segmento de clientes que presenta una mejor recuperación:



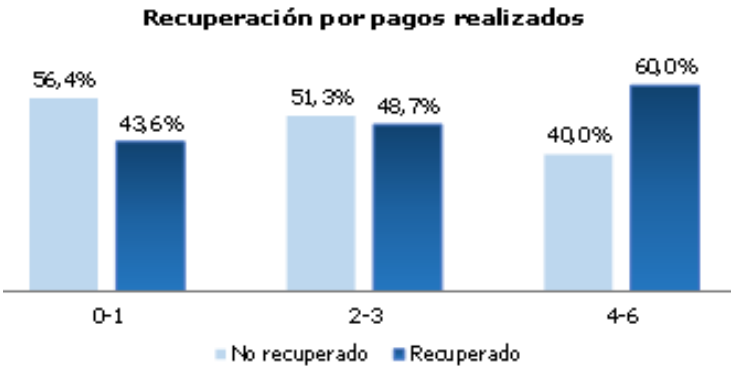
Elaboración propia

Se puede observar que el género femenino presenta un mejor comportamiento en el pago de la obligación comparado con el género masculino, pasando del 48 % al 52,5 %.

En cuanto a las garantías cuando el cliente no cuenta con ninguna de ellas por ejemplo el FNG su recuperación incrementa en un 15 %, alcanzando niveles del 58 %.

En el sector agropecuario como se mencionaba anteriormente es un sector que viene en crecimiento de acuerdo a las medidas implementadas por el gobierno, el mismo presenta un indicador de recuperación del 47,4 % (5 % por debajo de los clientes que no pertenecen a dicho sector)

Respecto a la antigüedad los clientes que pertenecen a la entidad financiera por más de 9 meses presentan un comportamiento de pago positivo alcanzando niveles del 61 %.



La recuperación por la mora del cliente en el momento de ejecutar el modelo y por pagos realizados son variables que se comportan de acuerdo a lo esperado, es decir a medida que la mora incrementa, la recuperación disminuye y a medida que los clientes realizan una mayor cantidad de pagos en los 6 meses anteriores la recuperación incrementa.

Capítulo 6

Conclusiones y Recomendaciones

- Del conjunto de modelos realizados se estableció como mejor opción el que proporcionó el menor porcentaje de error, es decir, el realizado por medio de árboles de decisión, con un 11,3 % el cual permite contar con un ajuste adecuado, demostrando la efectividad de este tipo de modelos en el sector microcreditico.
- Las variables que más afectan el incremento de la mora son la garantía, el genero y la mora en el momento de genera el modelo.
- Al evaluar el perfil del cliente, se identifica diferencias en la recuperación por género, sector agropecuario, garantía y antigüedad del cliente, lo cual es de vital importancia para la entidad dado que se pueden focalizar estrategias teniendo en cuenta estas variables.
- Ampliar el modelo a la mora mayor a 90 días con el fin de identificar la efectividad en estas franjas de difícil recuperación
- Establecer estrategias focalizadas en términos de recuperación teniendo en cuenta que el modelo presenta un ajuste adecuado.

Bibliografía

[1] Superintendencia Financiera de Colombia. *Calidad de cartera establecimientos de crédito*. URL: https://www.superfinanciera.gov.co/inicio/10082252.

[2] Horacio Fernandez Castro. *El modelo logístico: Una herramienta estadística para evaluar el riesgo de crédito*. Citado en el 2015. URL: http://www.redalyc.org/html/750/75040605/.

[3] Víctor Julio Zúñiga Vargas. *la Ejecución de un modelo de implementación de un Scoring estadístico para las micro finanzas*. Citado en el 2015. URL: http://dspace.unitru.edu.pe/handle/UNITRU/8501.

[4] Blanca Yadira Riera Naranjo. *Desarrollo de un modelo de scoring de segmentación de cobranzas para tarjeta de crédito de la banca de personas de Produbanco*. Citado en el 2018. URL: http://hdl.handle.net/10644/6213.

[5] Asobancaria. *Que es el microcredito*. Citado en Diciembre del 2011. URL: http://www.asobancaria.com/sabermassermas/que-es-el-microcredito/.

[6] Experian. *gestión de riesgo crediciticio*. Citado en el año 2015. URL: http://www.experian.es/gestion-del-riesgo-crediticio/informacion-y-scoring-scoring-y-analisis-de-credito.html.

[7] Erica Taucher. *Bioestadistica*. Segunda Edición. Editorial universitaria S.A, 1999.

[8] Universidad Nacional Mayor Marcos. *Modelo de regresión logística*. Citado en el año 2013. URL: http://sisbib.unmsm.edu.pe/bibvirtualdata/Tesis/Basic/Salcedo_pc/enPDF/Cap2.PDF.

[9] Universidad Autónoma de Manizales. *Regresión con variable dependiente binaria*. URL: http://www.uam.es/personal_pdi/economicas/rsmanga/docs/tema5_3_eco1.pdf.

[10] Damián Jorge Matich. *Redes Neuronales: Conceptos Básicos y Aplicaciones*. Citado en el año 2001. URL: ftp://decsai.ugr.es/pub/usuarios/castro/Material-Redes-Neuronales/Libros/matich-redesneuronales.pdf.

[11] IBM Knowledge Center. *El modelo de redes neuronales*. URL: https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/components/neuralnet/neuralnet_model.html.

[12] fernando Sancho Caparrini. *Redes Neuronales*. Citado en el año 2017. URL: http://www.cs.us.es/~fsancho/?e=72.

[13] Universidad Politécnica de Cartagena. *Introducción a los Árboles de Decisión*. URL: http://www.dmae.upct.es/~mcruiz/Telem06/Teoria/arbol_decision.pdf.

[14] Universidad Católica Andres Bello. *El proceso de la toma de decisiones*. URL: https://sites.google.com/site/gpsguayana/contenido/capitulo-iv---desarrollo-del-plan-para-proyectos/el-proceso-de-la-toma-de-decisiones.

[15] IBM Knowledge Center. *Modelos de árboles de decisión*. URL: https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/nodes_treebuilding.html.